ARTICLE

Estimating Genetic Effects and Quantifying Missing Heritability Explained by Identified Rare-Variant Associations

Dajiang J. Liu[1,2,3] and Suzanne M. Leal[1,2,*]

Next-generation sequencing has led to many complex-trait rare-variant (RV) association studies. Although single-variant association analysis can be performed, it is grossly underpowered. Therefore, researchers have developed many RV association tests that aggregate multiple variant sites across a genetic region (e.g., gene), and test for the association between the trait and the aggregated genotype. After these aggregate tests detect an association, it is only possible to estimate the average genetic effect for a group of RVs. As a result of the "winner's curse," such an estimate can be biased. Although for common variants one can obtain unbiased estimates of genetic parameters by analyzing a replication sample, for RVs it is desirable to obtain unbiased genetic estimates for the study where the association is identified. This is because there can be substantial heterogeneity of RV sites and frequencies even among closely related populations. In order to obtain an unbiased estimate for aggregated RV analysis, we developed bootstrap-sample-split algorithms to reduce the bias of the winner's curse. The unbiased estimates are greatly important for understanding the population-specific contribution of RVs to the heritability of complex traits. We also demonstrate both theoretically and via simulations that for aggregate RV analysis the genetic variance for a gene or region will always be underestimated, sometimes substantially, because of the presence of noncausal variants or because of the presence of causal variants with effects of different magnitudes or directions. Therefore, even if RVs play a major role in the complex-trait etiologies, a portion of the heritability will remain missing, and the contribution of RVs to the complex-trait etiologies will be underestimated.

Introduction

Nearly a decade of genome-wide association studies (GWASs) has led to the identification of many complex-trait associations with common variants (minor allele frequency [MAF] $>$ 5%).[1] The genetic-effect sizes of these common variants are for the most part very modest. Even for diseases with a strong genetic component, the identified common variants usually only explain a small portion of the total genetic heritability. For example, in a recent GWAS of human height, $>$100 significant single nucleotide polymorphism (SNP) markers were identified, but these collectively explained only ~10% of the heritability.[1] In a study of Crohn disease, $>$30 loci were identified, but they explain $<$10% of the overall heritability.[2,3] For GWASs, indirect mapping is performed, and therefore the proportion of heritability a particular causal variant contributes can be underestimated as a result of the incomplete linkage disequilibrium (LD) between the markers and the causal variant. Although genotyping larger sample sizes might elucidate additional loci for common variants with smaller effect sizes, it is clear that a major portion of the missing heritability remains unexplained.

Multiple hypotheses on missing heritability have been proposed. For example, it was suggested that the unexplained portion of genetic variance can be due to gene × gene interactions, gene × environment interactions, structural variation, epigenetics, and rare variants.[4] In particular, researchers are vigorously investigating the "common disease, rare variant" (CD/RV) hypothesis to determine whether rare variants (with MAF $\leq$ 1%) and low-frequency variants (with MAF between 1% and 5%) explain a large portion of the missing heritability.[5,6] There is solid evidence supporting the CD/RV hypothesis, which proposes that low-frequency and rare variants are involved in the etiology of complex traits.[6,7] Indirect association mapping approaches used in GWASs are underpowered to detect associations with rare variants because of the low LD (r^2) between common tagSNPs and rare causative variants.[8] Therefore, it is likely that many rare causative variants remain undetected even after extensive GWAS efforts. With the implementation of next-generation sequencing technology, large-scale sequence-based studies have been made possible. The validity of the CD/RV hypothesis and the proportion of missing heritability that is attributable to rare variants can now be examined.

For association mapping of rare variants, it is neither powerful nor numerically stable to analyze each variant individually. Gene-based tests, in which multiple rare variants in a gene region are jointly analyzed so that signals are aggregated and the number of tests is reduced, are usually performed. Many statistical tests have been proposed for detecting binary or quantitative-trait (QT) associations. Such tests include the combined multivariate and collapsing test (CMC),[8] gene- or region-based analysis of variants of intermediate and low frequency test (GRANVIL),[9]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Statistics, Rice University, Houston, TX 77005, USA
[3]Present address: Center of Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
*Correspondence: sleal@bcm.edu
http://dx.doi.org/10.1016/j.ajhg.2012.08.008.

weighted sum statistic test (WSS),[10] the variable threshold test (VT),[11] the kernel-based adaptive cluster test (KBAC),[12] the data adaptive sum test,[13] the RARECOVER method,[14] C-alpha test,[15] the sequence kernel association test (SKAT),[16] and the replication-based test (RBT), among others.[17] Among the rare-variant association methods, the combined multivariate and collapsing method, GRANVIL, WSS, KBAC, and RBT are based upon analyzing selected variants according to functional annotations or fixed cutoffs for the MAF. Variants at different nucleotide sites are weighted and aggregated. The multi-site genotype for the gene locus is coded as a single variable and tested for associations with the QT of interest. For these tests, applying different weighting schemes can increase power, in that the variants that are more likely to be causal can be assigned larger weights than the variants that are less likely to be causal. RARECOVER and VT are variable selection based methods. In addition to testing for associations with the QT of interest, RARECOVER and VT can also select the set of variants where the association statistics are maximized. C-alpha and SKAT differ from other methods in that they are based on the random-effects model. They assume a common distribution for the phenotypic effects of variants at different sites and test for the null hypothesis that the distribution has zero variation.

After a significant association is identified, the association signal needs to be interpreted, and relevant genetic parameters should be estimated. In particular, it is of interest to estimate the genetic variance explained by the variants that are jointly analyzed. When variable-selection-based methods are used, it is necessary to estimate the genetic variance explained by the set of selected variants that maximize the association test statistics (e.g., Z score statistics). The estimated genetic parameters are important for interpreting the association signal, quantifying the amount of heritability a specific gene contributes to the trait, making risk predictions, and designing replication studies.[18]

Most of the gene- and region-based rare-variant-association methods focus on testing the null hypothesis of no gene-QT associations, but they cannot be applied to estimating genetic parameters. For the rare-variant association tests that are based upon weighting or collapsing variants, if the weights are only dependent on the multi-site genotype, average genetic effects (AGEs), defined by the change in the QT per unit of change in the locus-specific genotype coding, can be estimated. Variable selection-based methods, e.g., VT and RARECOVER, calculate either CMC or GRANVIL statistics for different groups of rare variants and use their maximum as the test statistic. Therefore, after an association is detected, genetic effects can be estimated for the subset of variants where the CMC or GRANVIL statistics are maximized.

We prove theoretically that the AGEs for a group of rare variants can be efficiently estimated. The maximum-likelihood estimates or least-square estimates of AGEs obtained from an independent sample are consistent even if there are noncausal variants and/or if there are causal variants with heterogeneous effects in the gene region. We demonstrate theoretically that when multiple rare variants are jointly analyzed, the locus-specific genetic variance will always be underestimated, unless either all the variants that are jointly analyzed are causal and have equal effects or optimal weights can be assigned.

There can be considerable heterogeneity in rare-variant sites and frequencies even between closely related populations, e.g., neighboring European populations.[19] The genetic effects of rare variants in these populations can also be vastly different. It is therefore desirable to estimate the genetic effect by using the same sample from which the association is identified. However, a problem in doing so is that the naive estimates can be seriously inflated as a result of the winner's curse.[20–25] Additionally, if the AGE is estimated for the set of variants where the association test statistics (e.g., Z score statistics) are maximized, an even larger bias can be caused by the model selection procedure. We modified the resampling-based approach by Sun and Bull[25] and proposed appropriate bootstrap-sampling-split algorithms that can correct for the bias of AGE estimates. The algorithm is generic and can be applied to estimating genetic parameters for the associations identified by any rare-variant test.

The properties of the estimators of AGEs and the locus-specific genetic variance were investigated. Genetic data were generated according to a rigorous population-genetic model as described in Kryukov et al.[26] QTs were simulated with parameters estimated for clinically relevant complex traits.[11,27,28] In the simulation experiments, genetic-effect estimates are shown when association testing was performed with CMC,[8] VT,[11] and extended WSS[10,29] tests.

As an application, a published data set from the Dallas Heart Study was revisited.[27,28] Sequence data from *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910), *ANGPTL5* (MIM 607666), and *ANGPTL6* (MIM 609336) were tested for associations with nine metabolic QTs. For each QT, by implementing the CMC, VT, and WSS tests, we performed association analyses by using the 1,045 European American study subjects. The results coincide with those of previous studies of complex traits[30] and provide solid support for our simulation experiments.

Material and Methods

Genetic Models and Genetic Parameters of Interest

It is assumed that a set of S variant sites are jointly analyzed in the sample. S can be determined by prespecified frequency thresholds, functional annotations, etc. The multi-site genotype for individual i is given by $\overrightarrow{X}_i = (X_i^1, X_i^2, \cdots, X_i^S)$, where X_i^s is an indicator of whether individual i contains rare variants at site s, (e.g., $X_i^s = 1$ when the individual is homozygous for the rare allele or heterozygous at site s). The MAF for variants at site s is denoted by p_s. Hardy-Weinberg equilibrium is assumed in the general

population, so the rare-variant carrier frequency at site s satisfies $q_s = P(X_i^s = 1) = p_s^2 + 2p_s(1-p_s)$.

The following general QT model is assumed:

$$Y_i = \tilde{\alpha} + \sum_{s \in C} \tilde{\beta}_s X_i^s + \varepsilon_i \quad (1)$$

where $C \subseteq S$ is the set of causative variants that affect the quantitative phenotype. Rare-variant sites in S, but not in C, are noncausal and do not influence the QT. The error terms ε_i are assumed to be independently and identically distributed, i.e., $\varepsilon_i \sim N(0, \tau^2)$. Y_i can be either the QT of interest or the QT residual after an adjustment for confounders, such as age, sex, and population substructure.

When an association is identified in the stage 1 sample, it is of interest to estimate the following two parameters for a certain group G of rare variants: (1) the causative-variant effects, $\tilde{\beta}_s, s \in G \cap C$ and (2) the genetic variance explained by a group of rare variants G,

$$\sigma_G^2 = \mathrm{var}\Big(\sum_{s \in G \cap C} \tilde{\beta}_s X_i^s\Big) = \sum_{s \in G \cap C} (\tilde{\beta}_s)^2 \mathrm{var}(X_i^s) = \sum_{s \in G \cap C} (\tilde{\beta}_s)^2 q_s (1 - q_s) \quad (2)$$

The proportions of QT variance attributable to causative variants can be defined by

$$h_G^2 = \frac{\sigma_G^2}{\tau^2 + \sigma_G^2}. \quad (3)$$

If the overall heritability h^2 for the QT of interest is known, the proportion of heritability that is attributable to the identified group of variants can also be defined, i.e.,

$$F_G = \frac{h_G^2}{h^2}. \quad (4)$$

In practice, it is not possible to directly estimate $\tilde{\beta}_s$, σ_G^2, h_G^2, or F_G because causal and noncausal variants cannot be distinguished in real applications. Although in principle one can obtain unbiased estimates for $\tilde{\beta}_s$ by fitting a multiple regression model using all variants as covariates, it is neither numerically stable nor statistically efficient to analyze each rare variant individually.

Efficient Maximum-Likelihood Estimators for the Locus-Specific AGE

Instead of estimating genetic effects for each individual rare variant, we can estimate the AGE for a group of rare variants. The following model is usually used in the inferences and estimations of rare-variant genetic effects, i.e.,

$$Y_i = \alpha + \beta_{AGE} K(\vec{X}_i, Y_i) + e_i, \quad (5)$$

where the error term is assumed to follow a normal distribution $e_i \sim N(0, \tau^2)$ under the null hypothesis. The coding function $K(\vec{X}_i, Y_i)$ for the multisite genotype is generic and can incorporate many variant collapsing and grouping schemes. For instance, in the collapsing method, the coding function has the form of $K(\vec{X}_i, Y_i) = \delta(\sum_{s \in S} X_i^s > 0)$, where variant carriers are coded as 1 and noncarriers are coded as 0. For the WSS method that was first described by Madsen and Browning[10] and extended by Lin and Tang,[29] the coding function has the form of $K(\vec{X}_i, Y_i) = \sum_{s \in S} \hat{w}_s X_i^s$, and the weights depend on the multisite genotypes. Other types of weights are also possible. The parameter can be interpreted as the change of the QT per unit of change in the locus-specific genotype coding, i.e., $\beta_{AGE} = \partial E(Y_i | K(\vec{X}_i)) / \partial K(\vec{X}_i)$. The AGE-based genetic variance can also be estimated, i.e., $\sigma_{AGE}^2 = \beta_{AGE}^2 \mathrm{var}(K(\vec{X}_i))$.

It should be noted that for model (5), e_i is only normally distributed under the null hypothesis of no QT-gene association. Under the alternative hypothesis, because there are noncausative variants or because there are causal variants with effects in different magnitudes or directions, the residual errors may not always follow a normal distribution. However, as we showed in the supplemental methods, when the CMC coding is used, the model in formula (5) is still approximately correct. The locus AGE estimates obtained in an independent sample are asymptotically efficient and satisfy

$$\hat{\beta}_{AGE}^{CMC} \xrightarrow{a.s.} \beta_{AGE}^{CMC} = \frac{\sum_{s \in G \cap C} \tilde{\beta}_s q_s}{\sum_{s \in G} q^s}. \quad (6)$$

When weighted sum coding is used and the weights depend on the multisite genotypes, it is difficult to specify the joint-likelihood model of genotype and phenotypes. Instead, if a random population sample is collected, least-square estimates, which require less stringent distributional assumptions, can be obtained. In this case, as we show in the Supplemental Data available online, $\hat{\beta}_{AGE}^{WSS} \xrightarrow{a.s} \beta_{AGE}^{WSS} = \sum_s w_s \mathrm{cov}(X_i^s, Y_i) / \sum_s (w_s)^2 \mathrm{var}(X_i^s)$, where $\hat{w}_s \xrightarrow{a.s.} w_s$.

Because of noncausal variants or causal variants with effects of different magnitudes and directions, as well as the weights that are assigned to each variant site, the AGEs might be different in value from the causal variant effects, i.e., $\tilde{\beta}_s, s \in C$. However, the locus-specific genetic variance estimated from model (5) is always no greater than the true genetic variance, i.e., $\sigma_G^2 \leq \sigma_{AGE}^2$. Therefore, when multiple variants are jointly analyzed, the locus-specific genetic variance will be underestimated, and the estimates should be interpreted as a lower bound for the true locus-specific genetic variance.

Correcting for the Bias Due to the Winner's Curse

Most of the current sequence-based genetic studies can be underpowered because of the size of the data sets and the moderate effect sizes of variants involved in complex-trait etiologies. Therefore, the bias of the naive estimator can be substantial as a result of the winner's curse. Many methods have been developed to reduce the bias due to the winner's curse for single-marker tests in association studies of common variants; such tests include the likelihood-based methods by Xiao and Boehnke[23,24] and by Zöllner and Pritchard,[22] the resampling-based method by Sun and Bull,[25] and the Bayesian method by Xu et al.[31] Both the likelihood-based and Bayesian methods require evaluating power functions under the alternative hypothesis. However, this is impossible for rare-variant association studies because (1) p values for most rare-variant tests have to be obtained empirically via permutations, so there are no analytic formulas for calculating power and (2) the power for a rare-variant test depends on high-dimensional parameters, which include variant-site frequency spectrums and the phenotypic effects for different causal variant sites. Evaluating power functions on a fine grid in high-dimensional-parameter space is computationally intractable.

The resampling-based methods have desirable properties, in that they are nonparametric and do not require that power be calculated for a rare-variant test under the alternative hypothesis. However, the method by Sun and Bull[25] is not directly applicable to correcting the bias of the winner's curse in rare-variant

association analysis. This is for two reasons. First, in Sun and Bull,[25] the standardized genetic-effect estimates are used as the test statistic, i.e., $T_{\hat{\beta}} = \hat{\beta}/se(\hat{\beta})$. In rare variant analysis, a different test statistic, T, may be used for detecting associations. The correction procedure for the winner's curse can be complicated by the requirement that both T and $T_{\hat{\beta}}$ be statistically significant. Second, for some variable-selection-based tests, the set of variants that maximize the association test statistics are selected, and AGE parameters are estimated for this selected subset of variants. The bias of the estimates can be affected by both the winner's curse and the variable selection procedure.

To overcome these problems specific to rare-variant association testing, we developed two different bootstrap-sample-split algorithms. If the association was identified by a test that analyzes variants determined by some pre-specified criteria (e.g., fixed MAF cutoffs or functional annotations), genetic parameters for the set of variants that are jointly analyzed can be estimated via algorithm I. If the association is identified by a variable-selection-based method, such as VT or RARECOVER, genetic parameters for the set of selected variants that maximize the test statistics can be estimated via algorithm II.

In the BSS algorithm, for each bootstrap the original data set is split into a bootstrap sample and a residual sample, which are mutually exclusive. Hypothesis testing is performed with the bootstrap sample. If the statistic is significant, AGEs are estimated for both the bootstrap and the residual samples. When the two estimates are compared, the bias that is due to winner's curse and variable selection procedure can be quantified. So that stable estimates of the biases are obtained, the bootstrapping procedure is repeated multiple times. Deducting the bias from the naive estimates gives the BSS-corrected estimates. Technical details for the two algorithms can be found in the Supplemental Data.

Simulation of Genetic and QT Data and Association Analyses

Genetic data were generated according to Kryukov et al.[26] A conventional four-parameter model was used for describing the demographic history of the European population.[32] Purifying selection was also modeled for new nonsynonymous mutations. Details on this population-genetic model can be found in Kryukov et al.[26]

QT were simulated according to model (1). Different proportions of nucleotide sites were randomly chosen to be causative (i.e., 10%, 50%, and 90%), which covers a broad class of scenarios. The magnitudes of the genetic effects of causal variants are assumed to be inversely correlated with the MAFs, i.e,

$$|\tilde{\beta}_s| = \frac{\tilde{\beta}_{\max} - (\tilde{\beta}_{\max} - \tilde{\beta}_{\min})}{(\max_{s\in C}(p_s) - \min_{s\in C}(p_s))} \times (p_s - \min(p_s)), s\in C$$

A special case is when the magnitudes of the effects of the causal variants are equal, i.e., $\tilde{\beta}_{\min} = \tilde{\beta}_{\max}$. We considered scenarios where (1) $\tilde{\beta}_{\min} = \tilde{\beta}_{\max} = 0.25$, (2) $\tilde{\beta}_{\min} = \tilde{\beta}_{\max} = 0.5$, and (3) $\tilde{\beta}_{\min} = 0.125$, $\tilde{\beta}_{\max} = 0.75$. Under each set of values of $\tilde{\beta}_{\min}$ and $\tilde{\beta}_{\max}$, we considered scenarios where (1) all effects of causal variants are unidirectional and (2) 80% of causal variants increase the mean QT value and the remaining 20% of causal variants decrease mean QT value.

Data sets were simulated for random-population-based studies. For each replicate, genetic data and QTs for 20,000 individuals were generated. For a medium-sized study, 3,000 individuals were randomly chosen from the pool and analyzed in stage 1 (initial study) by the CMC, WSS, and VT methods. The statistical significance was evaluated for $\alpha = 0.05$. For a large-scale whole-exome study, 10,000 individuals were randomly sampled and analyzed in stage 1 by the CMC method. For this example, an exome-wide significance level of $\alpha = 2.5 \times 10^{-6}$ is used.

If the association test statistic was significant in stage 1, a stage 2 (replication) sample of equivalent size was selected from the remaining individuals in the pool. For the rare-variant association analyses implementing CMC and WSS, variants with MAF $\leq$ 1% were analyzed. For VT, Z score statistics with CMC coding are computed for each frequency threshold, and their maximum is used as the test statistic. Because we were interested in detecting associations with rare variants and because common variants can be individually tested for associations, we only analyzed variants with MAF $\leq$ 5%. This is slightly different from the original version of VT in Price et al.,[11] whose analysis included all variants in the gene region. Permutation was used to obtain the p value empirically for WSS and VT. For each scenario, we generated 30,000 replicates to obtain a sufficient number of significant replicates even for studies with low power.

Analysis of Sequence Data Set of ANGPTL3, ANGPTL4, ANGPTL5, and ANGPTL6

The sequence data set for *ANGPTL3, ANGPTL4*, *ANGPTL5,* and *ANGPTL6* was generated by the Dallas Heart Study (DHS). The DHS sample was collected from Dallas County residents whose lipids and glucose metabolism had been characterized and recorded.[33,34] For the four genes that were sequenced, a total of 348 nucleotide sites of sequence variations were uncovered. Most of the uncovered variants were rare, and 86% of them had MAFs $\leq$ 1%.[27] Nine phenotypes were measured and tested for their associations with rare genetic variants, i.e., body mass index (BMI), diastolic blood pressure (DiasBP), systolic blood pressure (SysBP), total cholesterol level (TCL), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TG), very-low-density lipoprotein (VLDL), and glucose (Gluc).

Association analyses were carried out with European-American samples. The CMC, WSS, and VT tests were used for analysis of the data set. The CMC and WSS methods were used for analysis of rare variants with MAF $\leq$ 3%. For each identified association, β_{AGE} - and AGE-based genetic variances σ^2_{AGE} are estimated for variants with MAF $\leq$ 3%. For the VT test, variants with MAF $\leq$ 5% were analyzed, and genetic parameters were estimated for the set of selected variants that maximize the Z score statistics.[11]

Results

Summary Statistics of Simulated Data Sets

Summary statistics are shown for different simulation scenarios, which include power, the cumulative causal-variant frequencies, all variant frequencies, and the number of variant nucleotide sites observed. Summary statistics are shown for the cases when rare-variant association analysis was performed with the CMC (Table 1), WSS (Table S1), and VT (Table S2). Because of the small phenotypic effects of causal variants and the low aggregated variant frequencies, the power to detect an association with a sample of 3,000 individuals is generally insufficient (e.g., <80%). In many scenarios, the power of VT is higher than that of the WSS and CMC methods. Because of a fixed

Table 1. Summary Statistics for Analyses Performed with CMC

$\tilde{\beta}_{max}$	$\tilde{\beta}_{min}$	Percentage of Causal Variants	Power[a]	Cumulative Causal-Variant Frequencies[b]	Cumulative Rare-Variant Frequencies[b]	$\overline{\beta_{AGE}^{CMC}}$ [c]	Total Number of Variant Nucleotide Sites
Causal Variant Effects Are Unidirectional							
0.25	0.25	10%	0.058	0.003	0.017	0.035	38.498
0.25	0.25	50%	0.157	0.012	0.019	0.151	38.937
0.25	0.25	90%	0.33	0.019	0.02	0.23	39.222
0.5	0.5	10%	0.083	0.005	0.018	0.107	38.71
0.5	0.5	50%	0.377	0.012	0.019	0.304	39.251
0.5	0.5	90%	0.754	0.016	0.018	0.455	38.964
0.75	0.125	10%	0.091	0.004	0.018	0.128	38.946
0.75	0.125	50%	0.483	0.01	0.018	0.37	39.378
0.75	0.125	90%	0.863	0.015	0.017	0.579	39.192
Causal Variant Effects Are Bidirectional[d]							
0.25	0.25	10%	0.057	0.003	0.017	0.021	38.458
0.25	0.25	50%	0.114	0.012	0.02	0.109	38.891
0.25	0.25	90%	0.203	0.018	0.02	0.173	39.023
0.5	0.5	10%	0.077	0.004	0.019	0.068	38.529
0.5	0.5	50%	0.263	0.012	0.02	0.232	38.925
0.5	0.5	90%	0.495	0.017	0.019	0.342	38.868
0.75	0.125	10%	0.085	0.004	0.018	0.087	38.862
0.75	0.125	50%	0.314	0.011	0.018	0.285	39.186
0.75	0.125	90%	0.572	0.016	0.018	0.426	39.24

Different portions of variants were randomly chosen to be causal, i.e., 10%, 50%, and 90%. Causal-variant effects are assumed to be either unidirectional or bidirectional. The magnitude of causal-variant effects is assumed to be either constant or inversely correlated with variant MAFs. Only those variants with MAFs $\leq$ 1% are analyzed in aggregate across the genetic region. The power, cumulative carrier frequencies of causal and all variants within the region, AGE, and number of variant nucleotide sites are reported.

[a]The power is calculated from 30,000 replicates under a significance level of $\alpha = 0.05$.

[b]We calculated the cumulative causal- and all-variant carrier frequencies within the genetic region by averaging over all replicates with significant test statistics.

[c]AGE is calculated for variants with MAF $\leq$ 1% by Equation 5, and reported values of $\overline{\beta_{AGE}^{CMC}}$ are averages of the model parameter β_{AGE}^{CMC} over the replicates with significant CMC statistics.

[d]Among the causal variants, 80% increase the mean QT value, and the remaining 20% decrease the mean QT value.

variant frequency threshold (i.e., MAF $\leq$ 1%), either causal variants with higher frequencies can be excluded from the analysis or more frequent noncausal variants can be included. The power advantage of VT over CMC can be large when there is a high portion of noncausal variants within the analyzed genetic region. For example, when 50% of the causal variants have a constant effect with $\tilde{\beta}_s = 0.5,\ s \in C$, the power for CMC, WSS, and VT tests are, respectively, 37.7% (Table 1), 33.5% (Table S1), and 45.4% (Table S2). If 90% of the variants are causal, the power of VT (79.4%; Table S2) is only slightly higher than the power of CMC (75.4%; Table 1) or WSS (72.4%; Table S1). The results are also shown for the simulation study where 10,000 samples are sequenced and an exome-wide significance level of $\alpha = 2.5 \times 10^{-6}$ is used (Table S3). For some scenarios, e.g., when 10% of the variants are causal and have effect $\beta_s = 0.25,\ s \in C$, the power can be extremely low (0.2%) even when 10,000 samples are sequenced (Table S3).

In our comparisons, for each replicate we generated a different pool of samples by using forward-time simulation. Therefore, the variant MAFs, the set of causal variants, and the value of the model parameter β_{AGE} can be different between replicates. It is important to note that the reported values $\overline{\beta_{AGE}^{CMC}}$, $\overline{\beta_{AGE}^{WSS}}$, and $\overline{\beta_{AGE}^{VT}}$ are averages of the model parameters β_{AGE}^{CMC}, β_{AGE}^{WSS}, and β_{AGE}^{VT} over the replicates with significant test statistics. Therefore, they can also be affected by the power to detect an association. The power is generally higher if the gene region has a larger AGE. Therefore, the average AGE that is calculated from the replicates with significant test statistics will be greater than when the average is obtained from all replicates. For example, when $\tilde{\beta} = 0.5$ and 50% of the variants are causal, the value of $\overline{\beta_{AGE}^{CMC}}$ and $\overline{\beta_{AGE}^{VT}}$ would be equal to 0.25 if averaged over all replicates, regardless of whether the test statistic is significant or not. For the replicates where the test statistic is significant, the mean proportion of causal

variants is >50%, and the average of AGEs is given by $\overline{\beta_{AGE}^{CMC}} = 0.304$ (Table 1) and $\overline{\beta_{AGE}^{VT}} = 0.353$ (Table S2) for rare-variant association analyses performed with CMC and VT.

Results are also shown for cases when a gene region contains causal variants with bidirectional effects—that is, when 80% of the causal variants increase mean QT values and the remaining 20% decrease the mean QT values. It can be observed that the bidirectionality of the causal-variant effect reduces the power of all tests and also causes a reduction in the AGE values. For example, when $\tilde{\beta}_s = 0.5,\ s \in C$ and 90% of the variants are causal, if the effects of all variants are unidirectional, the powers for CMC, WSS, and VT are 75.4%, 72.4%, and 79.4%, respectively, and the values of $\overline{\beta_{AGE}^{CMC}}$, $\overline{\beta_{AGE}^{WSS}}$, and $\overline{\beta_{AGE}^{VT}}$ are 0.455, 0.279, and 0.479, respectively. However, when the variant effects are bidirectional, the power decreases to 49.5%, 40.7%, and 59.7%, and $\overline{\beta_{AGE}^{CMC}}$, $\overline{\beta_{AGE}^{WSS}}$, and $\overline{\beta_{AGE}^{VT}}$ reduce to 0.342, 0.189, and 0.357, respectively (Table 1; Tables S1 and S2). Similarly, when 90% of the variants are causal, the effect sizes of causal variants are variable, $\tilde{\beta}_{\max} = 0.75$, $\tilde{\beta}_{\min} = 0.125$, and the effects are unidirectional, the powers for CMC, WSS, and VT are 86.3%, 88.0%, and 89.4%, respectively, and the mean values for $\overline{\beta_{AGE}^{CMC}}$, $\overline{\beta_{AGE}^{WSS}}$, and $\overline{\beta_{AGE}^{VT}}$ are 0.579, 0.382, and 0.635, respectively. When the effects of causal variants are bidirectional, the powers for the three tests decrease to 57.2%, 56.1%, and 68.9%, respectively, and $\overline{\beta_{AGE}^{CMC}}$, $\overline{\beta_{AGE}^{WSS}}$, and $\overline{\beta_{AGE}^{VT}}$ become 0.426, 0.255, and 0.467, respectively (Table 1; Tables S1 and S2).

Estimates for the AGE and Locus-Specific Variance in Aggregate Rare-Variant Analysis

Biases and variances for different estimators of AGE-based genetic variances were examined for a broad variety of models. First, the theoretical properties of AGE-based genetic-variance estimators were verified via simulation experiments. Specifically, the AGE-based genetic variance, σ_{AGE}^2 (i.e., $(\sigma_{AGE}^{CMC})^2$, $(\sigma_{AGE}^{WSS})^2$, and $(\sigma_{AGE}^{VT})^2$ for analyses performed with CMC, WSS, and VT, respectively) is always no greater than the true genetic variance in all the scenarios that were examined. The difference between σ_G^2 and σ_{AGE}^2 is larger when there is a higher portion of noncausal variants or when there is a mixture of causal variants that have effects in different directions. In our simulation experiment, the average values of AGE-based genetic variance [i.e., $\overline{(\sigma_{AGE}^{CMC})^2}$, $\overline{(\sigma_{AGE}^{WSS})^2}$, and $\overline{(\sigma_{AGE}^{VT})^2}$] and true genetic variance (i.e., $\overline{\sigma_G^2}$) were compared for the replicates with significant test statistics. For example, for the analysis using CMC, if 90% of the variants are causal and all causal variants have an effect of $\tilde{\beta}_s = 0.5,\ s \in C$, the average AGE-based variance is equal to $\overline{(\sigma_{AGE}^{CMC})^2} = 0.368 \times 10^{-2}$ (Figure 1F). In this case, $\overline{(\sigma_{AGE}^{CMC})^2}$ is 8.9% lower than the average true genetic variance, $\overline{\sigma_G^2}$. It should be noted that an unbiased estimate of the genetic variance could be obtained if optimal weights were assigned or if all variants were causal with equal effects. However, if only 50% of the variants are causal, $(\sigma_{AGE}^{CMC})^2$ is considerably reduced, i.e., $\overline{(\sigma_{AGE}^{CMC})^2} = 0.196 \times 10^{-2}$, which is 33.1% lower than the true genetic variance (Figure 1E). When there is a mixture of variants with effects in different directions, the discrepancy between AGE-based genetic variance, σ_{AGE}^2, and the true genetic variance, σ_G^2, may be increased further. For example, if 90% of the causal variants have positive effect 0.5 and the other 20% have a negative effect of −0.5, the average value of AGE-based genetic variance is $\overline{(\sigma_{AGE}^{CMC})^2} = 0.252 \times 10^{-2}$, which is 41.3% lower than the average value for the true locus-specific genetic variance (Figure 2F). When the genetic effects of causal variants are variable with MAFs, the locus-specific genetic variance will also be underestimated. For example, when $\beta_{\max} = 0.75$, $\beta_{\min} = 0.125$, and 90% of the variants are causal and have unidirectional effects, the locus-specific genetic variance is underestimated by 14.8%. When causal variants have bidirectional effects, $\overline{(\sigma_{AGE}^{CMC})^2}$ is 48.6% lower than $\overline{\sigma_G^2}$. Similar results are observed when the analysis is performed with WSS and VT.

It should be noted that $\overline{(\sigma_{AGE}^{CMC})^2}$ (Figures 1 and 2) and $\overline{(\sigma_{AGE}^{WSS})^2}$ (Figures S1 and S2) represent the average value of the genetic variance explained by variants with MAF ≤ 1%. They are not directly comparable to $\overline{(\sigma_{AGE}^{VT})^2}$ (Figures S3 and S4), which equals the average value of the genetic variance explained by the set of variants where the Z score statistics are maximized. For example, when 90% of variants are causal and all causal variants have an effect of $\tilde{\beta} = 0.5$, $\overline{(\sigma_{AGE}^{VT})^2}$ is 5.0% smaller than $\overline{\sigma_G^2}$ (Figure S3F). There is a much greater reduction in $\overline{(\sigma_{AGE}^{CMC})^2}$ (8.9%, Figure 1F) when variants with MAF ≤ 1% are analyzed. It is interesting to note that in the same scenario, $\overline{(\sigma_{AGE}^{WSS})^2}$ is 49.6% smaller than $\overline{\sigma_G^2}$ (Figure S1F). This is because, when WSS is used, the assigned weights are not optimal, i.e., noncausal, low-frequency variants are up-weighted, and higher frequency, causal variants are down-weighted. WSS can thus underestimate the locus-specific genetic variance to a greater extent than methods, such as CMC, that treat each variant interchangeably.

For the estimators obtained from an independent stage 2 sample, i.e., $(\hat{\sigma}_{AGE}^{S2})^2$, the biases are very small in all scenarios (<0.01 × 10^{-2}), which verifies the asymptotic consistency of the estimators $(\hat{\sigma}_{AGE}^{CMC,S2})^2$, $(\hat{\sigma}_{AGE}^{WSS,S2})^2$, and $(\hat{\sigma}_{AGE}^{VT,S2})^2$.

However, the biases due to the winner's curse or variable selection procedures can be considerable when estimation is performed with the discovery sample. This is compatible with the observations from common-variant association

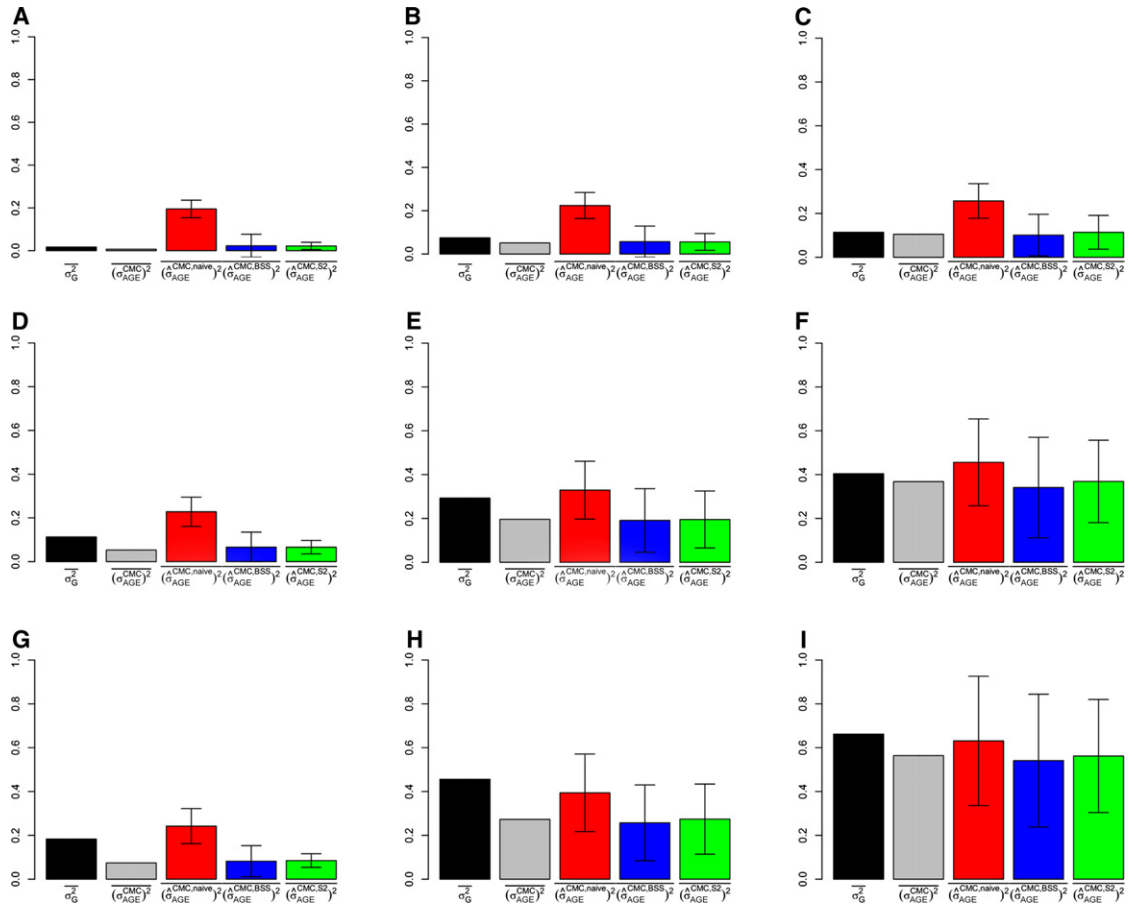

Figure 1. Estimates of Genetic Variance When Genetic Association Testing Is Performed with CMC
Data were generated under the assumption that the causal-variant effects are unidirectional. The replicates with significant test statistics were used for estimating genetic parameters for the variants with MAF $\leq$ 1%. Mean values and standard deviations are shown for the naive, BSS-corrected, and independent estimators, mean values are displayed as bar plots, and standard deviations are represented by error bars. The true genetic variance and AGE-based genetic variance were calculated analytically. The reported values $\overline{\sigma_G^2}$ and $\overline{(\sigma_{AGE}^{CMC})^2}$ are averages over the replicates with significant test statistics. Examined scenarios included those in which (A) 10% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (B) 50% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (C) 90% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (D) 10% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (E) 50% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (F) 90% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (G) 10% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$, (H) 50% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$, and (I) 90% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$.

analyses.[21,22,25] For example, if 50% of the variants are causal, and the causal-variant effect is $\tilde{\beta}_s = 0.5$, $s \in C$, the power of the CMC test is 37.7%. The average bias for the naive estimator $\overline{(\hat{\sigma}_{AGE}^{CMC,naive})^2} - \overline{(\sigma_{AGE}^{CMC})^2}$ is 0.133×10^{-2}, which is 67.8% of the average true parameter value (i.e., $\overline{(\sigma_{AGE}^{CMC})^2} = 0.196 \times 10^{-2}$) (Figure 1E). As a result of both the variable selection procedure and the winner's curse, the biases for the naive estimator can also be large when $(\hat{\sigma}_{AGE}^{VT,naive})^2$ is estimated for the group of variants where the Z score statistics are maximized. For instance, when analysis is performed with VT (Figure 2B) under the same scenario, the average bias $\overline{(\hat{\sigma}_{AGE}^{VT,naive})^2} - \overline{(\sigma_{AGE}^{VT})^2}$ is 0.153×10^{-2}, which is 35.6% more than the average true value ($\overline{(\sigma_{AGE}^{VT})^2} = 0.429 \times 10^{-2}$).

Finally, we examined the performances of BSS algorithms and showed that the biases can be consistently reduced. The biases of BSS-corrected estimators are comparable to those of independent estimators in most scenarios. For example, under the model where the causal-variant genetic effects are inversely correlated with MAFs with $\tilde{\beta}_{\max} = 0.75$ and $\tilde{\beta}_{\min} = 0.125$, if 10% of the variants are causal, the power for VT is 12.1%. The average biases for the BSS-corrected estimator $(\hat{\sigma}_{AGE}^{VT,BSS})^2$ is 0.016×10^{-2}, whereas the biases for the naive estimator and independent estimator are 0.211×10^{-2} and 0.011×10^{-2}, respectively. In certain scenarios where the power is higher, there can be some over-corrections for the BSS estimators. Examples include the scenario where $\tilde{\beta}_{\max} = 0.75$, $\tilde{\beta}_{\max} = 0.125$, and 90% of the variants are causal. In this case, the power for VT is 89.4%, and the average value of the estimate $(\hat{\sigma}_{AGE}^{VT,BSS})^2$ is deflated by -0.049×10^{-2} (Figure 1I). The performance for the BSS estimator is similar when analysis is performed by the VT or WSS method.

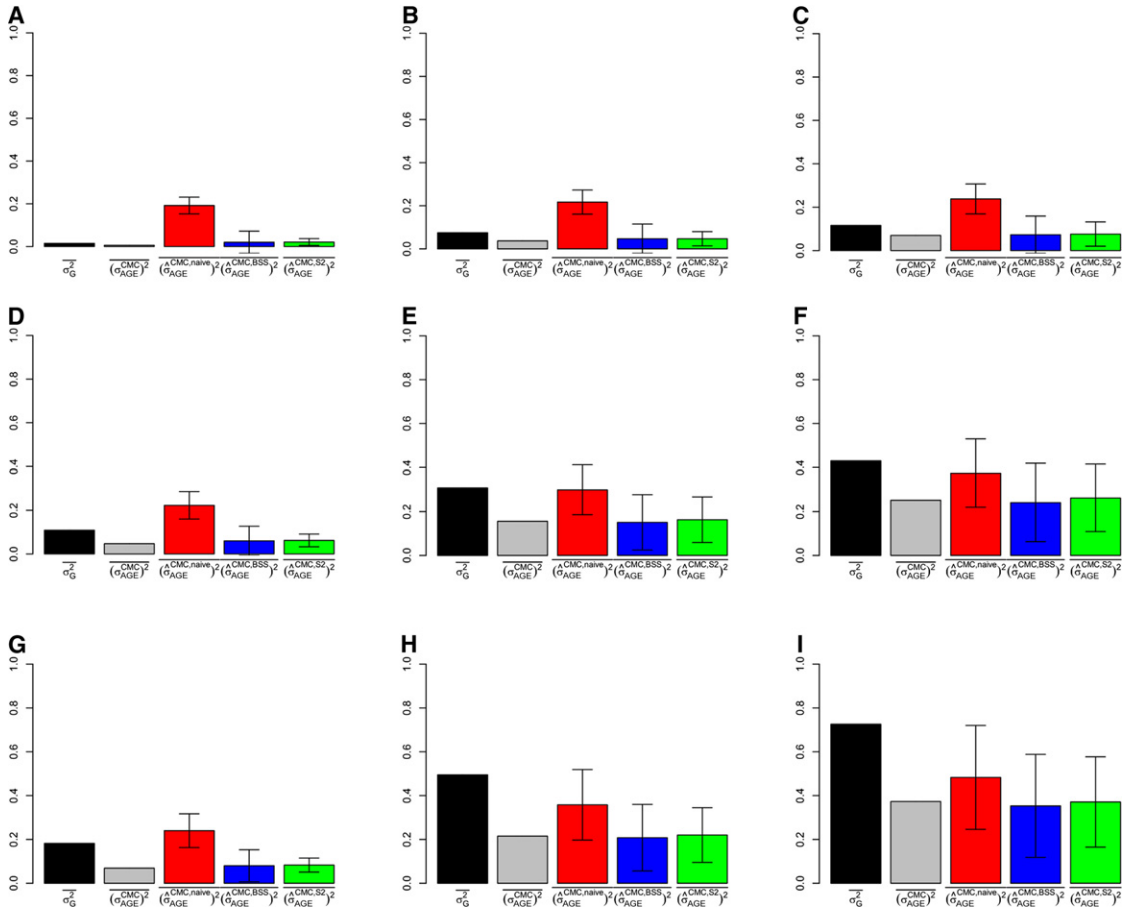

Figure 2. Estimates of Genetic Variance When Genetic Association Testing Is Performed with CMC
Data were generated under the assumption that the causal-variant effects were bidirectional. It is assumed that 80% of the causal variants increase the mean QT value, whereas the remaining 20% decrease the mean QT value. The replicates where the test statistic is significant were used for estimating genetic parameters for the variants with MAF $\leq$ 1%. Mean values and standard deviations are shown for the naive, BSS-corrected, and independent estimators, mean values are displayed as bar plots, and standard deviations are represented by error bars. The true genetic variance and AGE-based genetic variance were calculated analytically. The reported values $\overline{\sigma_G^2}$ and $\overline{(\sigma_{AGE}^{CMC})^2}$ are averages over the replicates with significant test statistics. Examined scenarios included those in which (A) 10% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (B) 50% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (C) 90% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.25$, (D) 10% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (E) 50% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (F) 90% of variants are causal and $\tilde{\beta}_{\max} = \tilde{\beta}_{\min} = 0.5$, (G) 10% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$, (H) 50% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$, and (I) 90% of variants are causal, $\tilde{\beta}_{\max} = 0.75$, and $\tilde{\beta}_{\max} = 0.125$.

We also compared the variances of different estimators. In fact, the standard deviations for the BSS-corrected estimators can be larger than those for the naive and independent estimators. This is because the BSS-corrected estimates are obtained from only a fraction of the sample (i.e., the residual sample for each bootstrap). For instance, in the analyses by CMC, when $\tilde{\beta}_s = 0.5$, $s \in C$ and 50% of the variants are causal, the standard deviation for the BSS-corrected estimator is 0.145×10^{-2}, which is slightly larger than that of the naive estimator (0.132×10^{-2}) and the independent estimator (0.130×10^{-2}) (Figure 1E).

Analysis of Sequence Data Set of ANGPTL3, ANGPTL4, ANGPTL5, and ANGPTL6

Association testing was performed with the CMC and VT methods. For each identified association, the naive and corrected estimators for β_{AGE} and σ_{AGE}^2 were reported. For the analyses using the CMC and WSS methods, rare variants with MAF $\leq$ 3% were grouped and jointly analyzed. Two significant associations, i.e., the association between *ANGPTL4* and TG ($p = 0.002$) and the association between *ANGPTL4* and VLDL ($p = 0.005$), were identified by CMC (Table 2). Only the association between *ANGPTL4* and TG was identified by WSS ($p = 0.038$). Association analyses were also performed with VT, where variants with frequency $\leq$ 5% were analyzed. Three associations, i.e., the associations between *ANGPTL4* and TG ($p = 0.005$ and MAF threshold = 0.014), between *ANGPTL4* and VLDL ($p = 0.014$ and MAF threshold = 0.014), and between *ANGPTL5* and TCL ($p = 0.008$ and MAF threshold = 0.031) were identified (Table 3).

We examined more closely the association between *ANGPTL5* and TCL; this association was only identified by the VT test. One variant (c.803C>T [p.Thr268Met];

Table 2. CMC and WSS Association Analyses for the *ANGPTL3*, *ANGPTL4*, *ANGPTL5*, and *ANGPTL6* Sequence Data Set

Gene	Trait	p Value	Naive Estimates		BSS-Corrected Estimates	
			$\widehat{\beta}_{AGE}^{navie}$	$(\widehat{\sigma}_{AGE}^{naive})^2$	$\widehat{\beta}_{AGE}^{BSS}$	$(\widehat{\sigma}_{AGE}^{BSS})^2$
Analysis by CMC						
ANGPTL4	TG	0.002	−0.476	1.067×10^{-2}	−0.426	0.817×10^{-2}
ANGPTL4	VLDL	0.005	−0.436	0.893×10^{-2}	−0.375	0.687×10^{-2}
Analysis by WSS						
ANGPTL4	TG	0.038	0.493	0.486 × 10−2	0.356	0.196×10^{-2}

Variants with MAFs ≤ 3% were analyzed. For each nominally significant association, β_{AGE} - and AGE-based genetic variance, σ_{AGE}^2, was reported.

RefSeq accession number NM_178127.4) in the region of interest has an MAF of ~0.031 and thus was not included in the analysis that used CMC and WSS. This variant on its own is nominally significantly associated with TCL ($p = 0.013$). Therefore, when this potential causal variant was excluded from the analysis because of the fixed MAF cutoff that was applied, CMC and WSS failed to detect the association signal.

For all the nominally significant associations, AGE and AGE-based variance were estimated via both the naive estimator and the BSS-corrected estimator. Concordant with our simulation experiment, the BSS-corrected estimator is usually smaller in scale, which indicates that the naive estimator can be inflated as a result of the winner's curse. For some associations, the difference can be fairly small. For example, for the association between *ANGPTL4* and TG, the estimated AGE and AGE-based variances for the set of variants with MAF ≤ 0.03 are, respectively, (1) $\widehat{\beta}_{AGE}^{CMC,naive} = -0.476$ and $(\widehat{\sigma}_{AGE}^{CMC,naive})^2 = 1.067 \times 10^{-2}$ and (2) $\widehat{\beta}_{AGE}^{CMC,BSS} = -0.426$ and $(\widehat{\sigma}_{AGE}^{CMC,BSS})^2 = 0.817 \times 10^{-2}$. However, for some associations, the difference between the two estimators can be large. For example, for the association between *ANGPTL5* and TCL, the naive estimators calculated for the group of selected variants (i.e., variants with MAF ≤ 0.031) are $\widehat{\beta}_{AGE}^{VT,naive} = 0.347$ and $(\widehat{\sigma}_{AGE}^{VT,naive})^2 = 0.753 \times 10^{-2}$, whereas the corrected estimators are $\widehat{\beta}_{AGE}^{VT,BSS} = 0.181$ and $(\widehat{\sigma}_{AGE}^{VT,BSS})^2 = 0.111 \times 10^{-2}$.

The estimates of locus-specific genetic variances obtained via WSS are smaller in scale than those obtained via CMC. Specifically, for the association between *ANGPTL4* and TG, the estimates are $(\widehat{\sigma}_{AGE}^{WSS,naive})^2 = 0.486 \times 10^{-2}$ and $(\widehat{\sigma}_{AGE}^{WSS,naive})^2 = 0.196 \times 10^{-2}$, indicating that the assigned weights are not optimal and might affect the estimation of locus-specific genetic variance in aggregate analysis.

For the association between TG and *ANGPTL4* that was replicated with an independent sample, the BSS-corrected estimate for AGE-based genetic variance (i.e., $(\widehat{\sigma}_{AGE}^{CMC,BSS})^2$) is equal to 0.817×10^{-2}. This suggests that rare variants in *ANGPTL4* explain at least ~0.8% of the overall phenotypic variance according to Equation 3. In a previous study, the overall heritability of TG was estimated to be 0.49.[35] Therefore, according to Equation 4, rare variants in *ANGPTL4* contribute ≥1.63% of the overall heritability of TG.

Discussion

In this article, the problem of estimating locus-specific genetic effects and variances was investigated for sequence-based genetic studies of rare variants. The results have important implications for interpreting the identified associations. For a given group of rare variants that are jointly analyzed, we showed that it is possible to estimate the AGE. The maximum-likelihood estimates or least-square estimates are asymptotically consistent even if the model used in the estimation differs from the true underlying genetic model. Estimates of the AGE can be affected by the presence of noncausal variants and/or causal variants with effects of different directions or magnitudes.

Table 3. VT Association Analyses for the *ANGPTL3, ANGPTL4, ANGPTL5*, and *ANGPTL6* Sequence Data Set

Gene	Trait	p Value	Naive Estimates		BSS-Corrected Estimates		MAF Cutoffs
			$\widehat{\beta}_{AGE}^{navie}$	$(\widehat{\sigma}_{AGE}^{naive})^2$	$\widehat{\beta}_{AGE}^{BSS}$	$(\widehat{\sigma}_{AGE}^{BSS})^2$	
ANGPTL4	TG	0.005	−0.476	1.067×10^{-2}	−0.444	0.758×10^{-2}	0.014
ANGPTL4	VLDL	0.014	−0.436	0.893×10^{-2}	−0.299	0.498×10^{-2}	0.014
ANGPTL5	TCL	0.008	0.347	0.753×10^{-2}	0.181	0.111×10^{-2}	0.031

For each nominally significant association, the MAF threshold where the *Z* score statistics are maximized is reported. β_{AGE} - and AGE-based genetic variance, σ_{AGE}^2, was reported for the set of variants determined by the optimal MAF threshold.

On the other hand, when multiple variants are jointly analyzed, the AGE-based genetic variance defined in the aggregate analysis is always no greater than the true locus-specific genetic variance under a broad variety of phenotypic models. Therefore, the locus-specific genetic variance and the proportion of missing heritability explained by the gene locus will be underestimated.

The estimates of locus-specific genetic variance in aggregate rare-variant association analysis will be affected by the presence of noncausal variants and/or the presence of causal variants with effects with different magnitudes or directions. With an increasing proportion of noncausal variants or a higher level of heterogeneities in the effects of causal variants, the locus-specific genetic variance can be underestimated to a greater extent. In addition, the estimates of locus-specific genetic effects can also be affected by the weights that are assigned to each variant site. When suboptimal weights are used (e.g., the weights used in Tang and Lin's extension[29]), the estimated value of σ^2_{AGE} can be much lower than the true locus-specific genetic variance and should be interpreted with caution.

When genetic model parameters are estimated from the stage 1 sample where the association is identified, the estimates can be inflated as a result of the winner's curse. The size of the bias due to the winner's curse was investigated for rare-variant association analysis. Compatible with observations in common-variant association analysis, the magnitude of the bias is usually inversely correlated with the power of the study. For some underpowered studies, the bias can be considerably large. Inflated estimates of genetic effects can lead to the underestimation of the size of samples that are needed for follow-up studies and hence the inability to replicate a genuine association. For association analysis of common variants, it is usually possible to obtain unbiased estimates of genetic parameters from a stage 2 replication sample. However, for rare variants, there is considerable heterogeneity of rare-variant sites and frequencies even for closely related populations,[19] which affects the estimates of genetic effects. It is crucial to be able to obtain unbiased estimates from the stage 1 study where the association is identified. Through the BSS algorithms that we developed, the bias due to the winner's curse can be consistently reduced even for studies with very low power (~10%). The algorithm can be used with all rare-variant tests that are based upon weighting or collapsing rare variants or on variable selection techniques.

The Dallas Heart Study data set was analyzed with the CMC, WSS, and VT tests. The identified associations coincide with previously published results. The association between *ANGPTL4* and TG is mainly driven by the p.GLU40LYS variant (c.118G>A; RefSeq accession number NM_020581.2), which is relatively common within the European population (the carrier frequency is ~3%). The association between the c.118G>A variant in *ANGPTL4* and TG was replicated in an independent data set. In our analysis using VT, c.118G>A was among the set of selected rare variants for which the *Z* score statistics are maximized. The estimate of β_{AGE} for the set of selected variants is −0.44 SD after correction, which is similar to the naive estimate. For the analysis using WSS, the assigned weights are inversely correlated to MAFs. Therefore, the c.118G>A variant that is potentially causal was assigned a smaller weight than other lower-frequency variants, which may not be causal. Concordant with our simulation experiment, it is clear that the locus-specific genetic variance can be underestimated to a greater extent if the assigned weights are not optimal. For the association between *ANGPTL5* and TCL, the BSS-corrected genetic effect estimates are considerably smaller than the naive estimates. This is an indication that the study might be underpowered or that the identified signal might be a false-positive result. Therefore, a replication study is needed to confirm the identified association.[36]

Approaches based on the random-effects model can also allow estimation of genetic effects or variance components. However, given the complexity of the model, it is hard to analytically explore the statistical properties of the estimator under the alternative hypothesis. It was well known that the inference can be biased if the assumptions of the random-effects model are violated.[37] Because of noncausal variants or the fact that lower-frequency variants are more likely to be functionally deleterious when effect sizes are large,[6] many key assumptions of the random-effects model can be violated. It is therefore unclear how to interpret the estimates obtained from these models.

The simulation experiments that are shown in this article were limited to population-based random-sampling designs. Because of the high sequencing cost, many studies of complex traits have sequenced selected samples with extreme traits. We have also simulated data sets for extreme-sampling studies and examined the properties of the AGE estimates. The results for the analysis using VT and CMC remain unchanged, and the methods are also applicable to extreme-sampling studies (data not shown). For the analysis by WSS, the estimates have to be obtained by least-square methods and therefore are not applicable for studies where samples with extreme quantitative traits are sequenced.

With the large-scale implementation of second-generation sequencing, the cost of generating and analyzing sequence data is expected to drop rapidly. The scale of sequence-based association studies will thus quickly expand. In large-scale sequencing studies with thousands of individuals, it is possible to analyze low-frequency variants (with MAF between 1% and 5%) or higher-frequency rare variants (e.g., with MAF between 0.5% and 1%) individually. In addition, with a large reference panel and cohort of samples, some low-frequency variants can also be accurately imputed from genotype data and analyzed individually. However, as a result of purifying selections and recent, rapid expansion of the human population, there are numerous variant sites with very low frequencies.[18] As a larger number of samples are sequenced, many more nucleotide sites with one or a few variants

will be uncovered within a data set.[26] In addition, some identified associations are clearly driven by multiple variants with very low frequencies, such as singletons. Given that the cumulative frequencies of these variants can be sufficiently large for detecting associations, it is still necessary and meaningful to aggregate rare variants in the gene region and analyze them collectively. Therefore, the methods introduced in this article will still be of great importance for future large-scale sequencing studies.

As sequence sample size increases, loci with smaller effect sizes can also be detected with high power. The AGE values for these new loci are expected to be smaller. For example, the identification of causal variants with smaller effects or the presence of a greater portion of noncausal variants in the genetic region can both lead to the decrease of the magnitude of the AGE estimates. When rare-variant association analysis shifts to the use of whole-genome sequence data, the unit of analysis will not be as clear as for exome sequencing (for which it is usually a gene). Additionally, the annotation for noncoding regions of the genome is not as straightforward as for coding regions. The genetic locus can thus be contaminated with a higher proportion of noncausal variants. The heritability that the genetic locus contributes can be underestimated to a greater extent in aggregate analyses.

By reducing the discrepancy between AGE-based genetic variance, σ^2_{AGE}, and true locus-specific genetic variance, σ^2_G, the estimation for a gene region's contribution to trait heritability will be improved. It is clear from our proof in the supplemental material that σ^2_{AGE} and σ^2_G are equivalent if all variants in the genetic region have homogeneous effect or when the optimal weights can be assigned. Therefore, one possible future direction of research is to integrate information from laboratory-based experiments or bioinformatics so that noncausal and causal variants can be more accurately distinguished and weighted or to use variable-selection-based methods to select variants with similar genetic effects. These new methods will be helpful for improving the estimates of locus-specific genetic variance. However, as we proved theoretically, when multiple rare variants are jointly analyzed, the locus-specific genetic variance will always be underestimated. The bias cannot be eliminated by any variable selection or weighting methods, and the estimated value of σ^2_{AGE} should be interpreted as a lower bound for the true locus-specific genetic variance.

With growing application of next-generation sequencing, many rare-variant associations will be detected. These findings need to be correctly interpreted. It is important to be able to estimate relevant genetic parameters of interest and quantify the proportion of heritability these novel associations explain. However, when aggregate analysis is performed for the detection of rare-variant associations, the inclusion of noncausal variants or the presence of causal variants with effects of different magnitudes or directions might cause the heritability contributed by rare variants to be underestimated.

Supplemental Data

Supplemental Data include Supplemental Methods, six figures, and three tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

This research is supported by National Institutes of Health grants MD005964 and HL102926 (S.M.L.). We would like to thank Jonathan Cohen and Helen Hobbs for providing us with data on the *ANGTPL* family genes from the Dallas Heart Study, which was supported by National Institutes of Health grant RL1HL092550 (J.C.). We would also like to thank Shamil Sunyaev (S.S.) for sharing the simulated genetic datasets from his projects, which were supported by National Institutes of Health grant MH084676 (S.S). Computation for this research was supported in part by the Shared University Grid at Rice, which was funded by the National Science Foundation under grant EIA-0216467, and by a partnership among Rice University, Sun Microsystems, and Sigma Solutions, Inc.

Received: February 16, 2012
Revised: June 19, 2012
Accepted: August 8, 2012
Published online: September 27, 2012

References

1. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
2. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
3. Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.
4. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
5. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
6. Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. *40*, 695–701.
7. Gibson, G. (2011). Rare and common variants: Twenty arguments. Nat. Rev. Genet. *13*, 135–145.
8. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
9. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.

10. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
11. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
12. Liu, D.J., and Leal, S.M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. *6*, e1001156.
13. Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. *70*, 42–54.
14. Bhatia, G., Bansal, V., Harismendy, O., Schork, N.J., Topol, E.J., Frazer, K., and Bafna, V. (2010). A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput. Biol. *6*, e1000954.
15. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.
16. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
17. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. *7*, e1001289.
18. Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. *1*, 131.
19. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.
20. Joiner, K.A. (2005). Avoiding the winner's curse in faculty recruitment. Am. J. Med. *118*, 1290–1294.
21. Zhong, H., and Prentice, R.L. (2010). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. Genet. Epidemiol. *34*, 78–91.
22. Zollner, S., and Pritchard, J.K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am. J. Hum. Genet. *80*, 605–615.
23. Xiao, R., and Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. Genet. Epidemiol. *33*, 453–462.
24. Xiao, R., and Boehnke, M. (2011). Quantifying and correcting for the winner's curse in quantitative-trait association studies. Genet. Epidemiol. *35*, 133–138.
25. Sun, L., and Bull, S.B. (2005). Reduction of selection bias in genomewide studies by resampling. Genet. Epidemiol. *28*, 352–367.
26. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876.
27. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. *39*, 513–516.
28. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J. Clin. Invest. *119*, 70–79.
29. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. *89*, 354–367.
30. Darvasi, A. (2006). Closing in on complex traits. Nat. Genet. *38*, 861–862.
31. Xu, L., Craiu, R.V., and Sun, L. (2011). Bayesian methods to overcome the winner's curse in genetic studies. Annals of Applied Statistics *5*, 201–231.
32. Adams, A.M., and Hudson, R.R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics *168*, 1699–1712.
33. Browning, J.D., Szczepaniak, L.S., Dobbins, R., Nuremberg, P., Horton, J.D., Cohen, J.C., Grundy, S.M., and Hobbs, H.H. (2004). Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology *40*, 1387–1395.
34. Victor, R.G., Haley, R.W., Willett, D.L., Peshock, R.M., Vaeth, P.C., Leonard, D., Basit, M., Cooper, R.S., Iannacchione, V.G., Visscher, W.A., et al.; Dallas Heart Study Investigators. (2004). The Dallas Heart Study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am. J. Cardiol. *93*, 1473–1480.
35. Middelberg, R.P., Martin, N.G., Montgomery, G.W., and Whitfield, J.B. (2006). Genome-wide linkage scan for loci influencing plasma triglycerides. Clin. Chim. Acta *374*, 87–92.
36. Liu, D.J., and Leal, S.M. (2010). Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am. J. Hum. Genet. *87*, 790–801.
37. Zhang, D., and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics *57*, 795–802.